

# Big Data et la santé

## Big Data, c'est quoi ?

- **Collecte, stockage et exploitation de masses de données**
  - Capturer de façon automatique et anonyme une très grande quantité d'informations, les traiter avec des algorithmes
  - si possible en temps réel (ex comparateurs de prix)
  - **Objectif** : modifier nos comportements !
  - **Origine** : le développement des réseaux et des objets connectés, la puissance et la rapidité des calculs, le Cloud computing (données stockées et accessibles en ligne)
  - **Une tentative de définition : critères des « 5V »**
    - Volume
    - Variété (données hétérogènes, diverses sources et différents formats)
    - Vélocité : accès et traitement temps réel
- + Véracité : précision et exactitude
- ++ **V**aleur : capacité à créer de la valeur

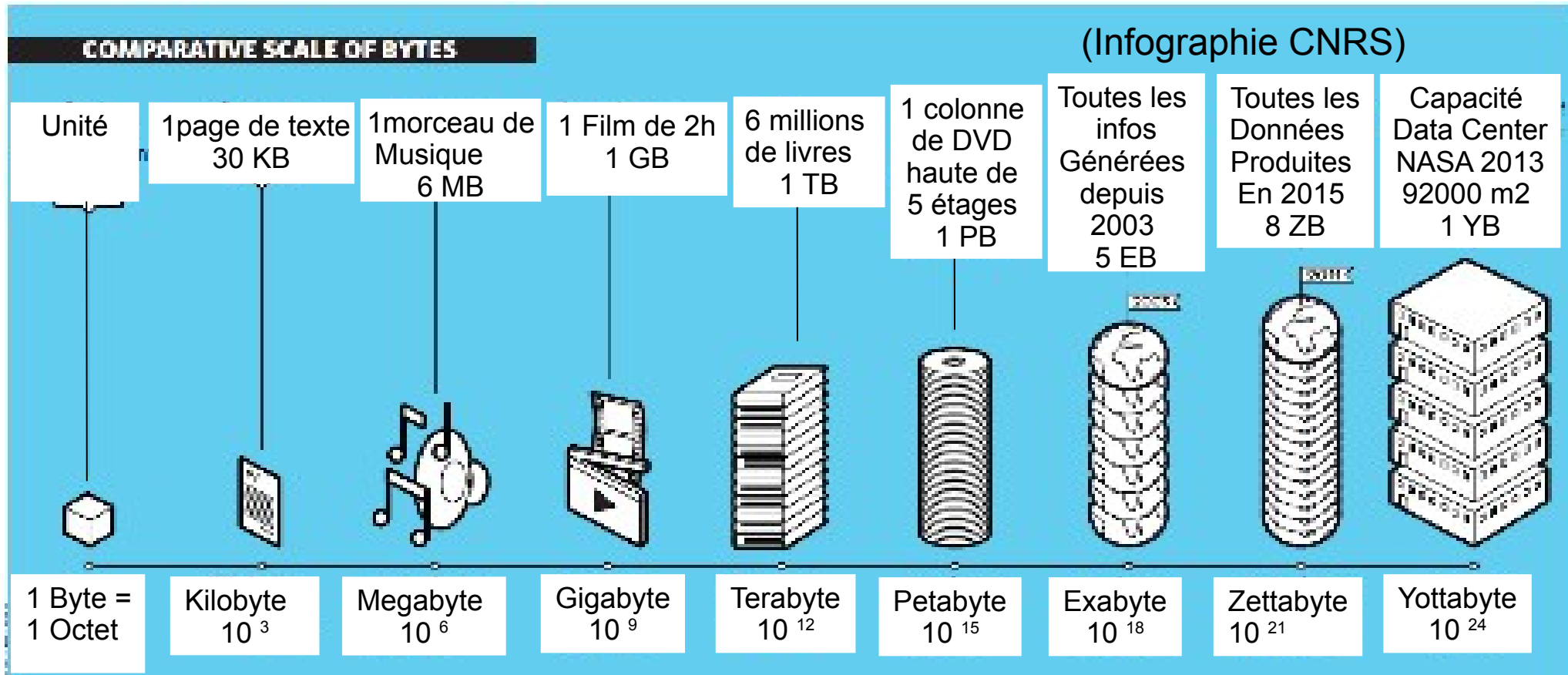
Utilisation de grappes de serveurs (clusters d'ordinateurs indépendants)

**Données => traitement (interactions, corrélations) => interprétation => prédiction**

8% du PIB européen en 2020 ?

80% d'entre elles sont détenues par les GAFA (Google, Amazon, Facebook, Apple, etc.).

# Un nombre de données gigantesque !

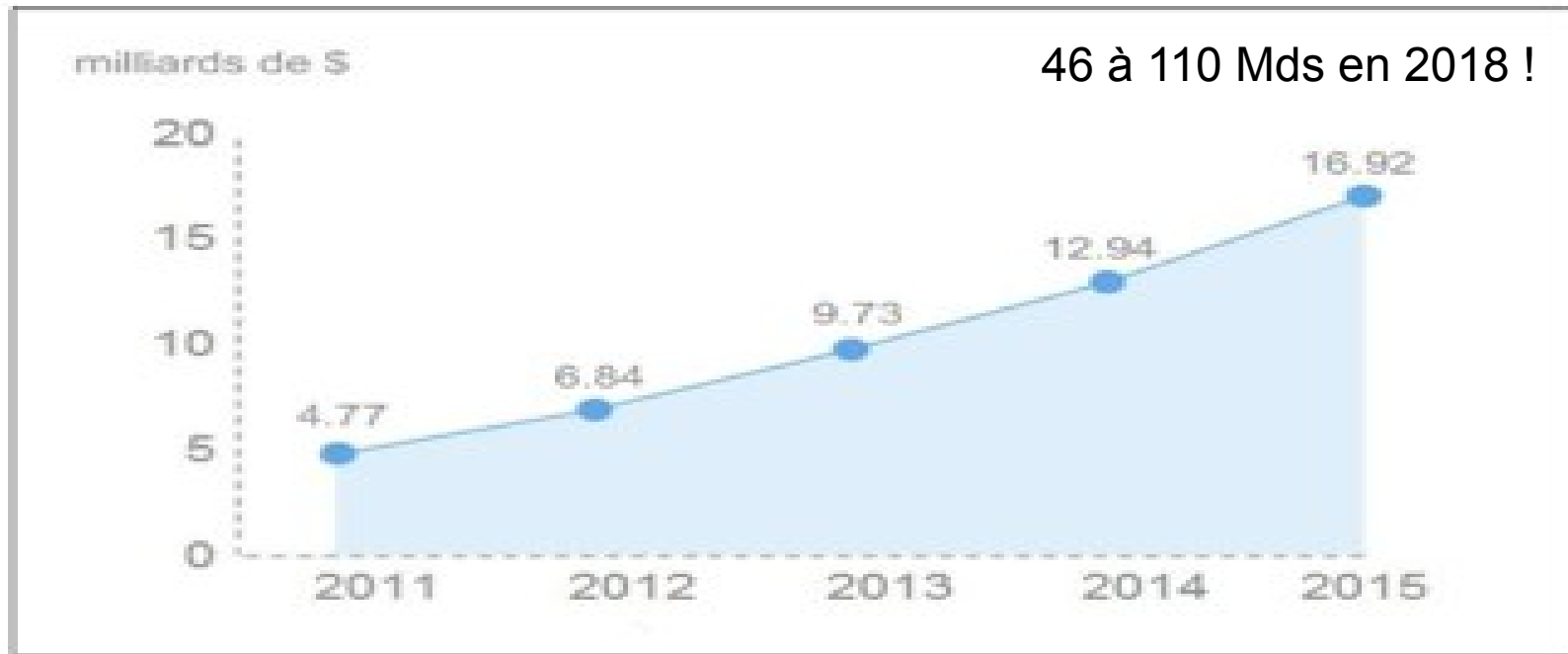


La plupart viennent d'Internet mais aussi de capteurs et des objets connectés

## Problèmes :

- le stockage => développement de mémoires non volatiles à l'échelle nano
- l'énergie pour refroidir tout cela => nouveaux matériaux, nouvelles architectures

# Un gigantesque marché



Estimation de l'évolution du marché des Big Data d'ici 2015

## Données

- Financières
- Recherche en physique (particules, astronomie)
- Santé (génomique, médecine)
- Marketing et de vente
- Gestion (villes, aéroports, infrastructures à risques)
- Météo
- Politiques, sécurité (le plus gros centre de stockage est celui de la NSA)

# Dans le domaine de la santé

Quelles données ?

## 1. Images et données médicales des patients

=> PACS : développement essentiel pour l'imagerie

=> e-santé : seulement 9% des hôpitaux européens permettent aux patients d'accéder à un dossier médical en ligne !

## 2. Recherche et études épidémiologiques

=> enquêtes, données de cohortes pour études internationales, évaluation des médicaments (pharmacovigilance= recherche de signaux faibles détectés dans les cohortes), santé publique et sécurité sanitaire (information des citoyens)

## 3. Génomique et biologie moléculaire

=> le coût du séquençement du génome a fortement baissé

=> détection de maladies génétiques, prédispositions

## 4. Données issues des Objets connectés

=> les vraies Big data à forte valeur ajoutée

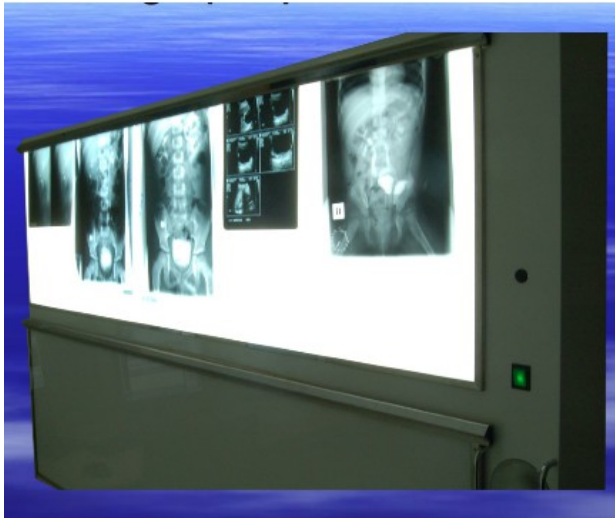
=> plutôt forme et bien-être (70% des applications) : auto-mesure de données physiques et biologiques et de l'activité physique

=> télésurveillance de maladies chroniques, avec l'objectif de 50 000 patients suivis en 2017 (75% des dépenses de santé dans les pays riches)

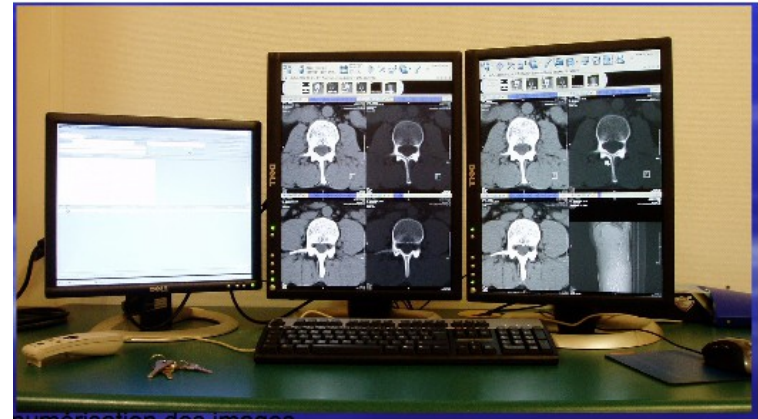
=> les risques !

# Imagerie : PACS ou Picture Archiving and Communication System

Avant : le film



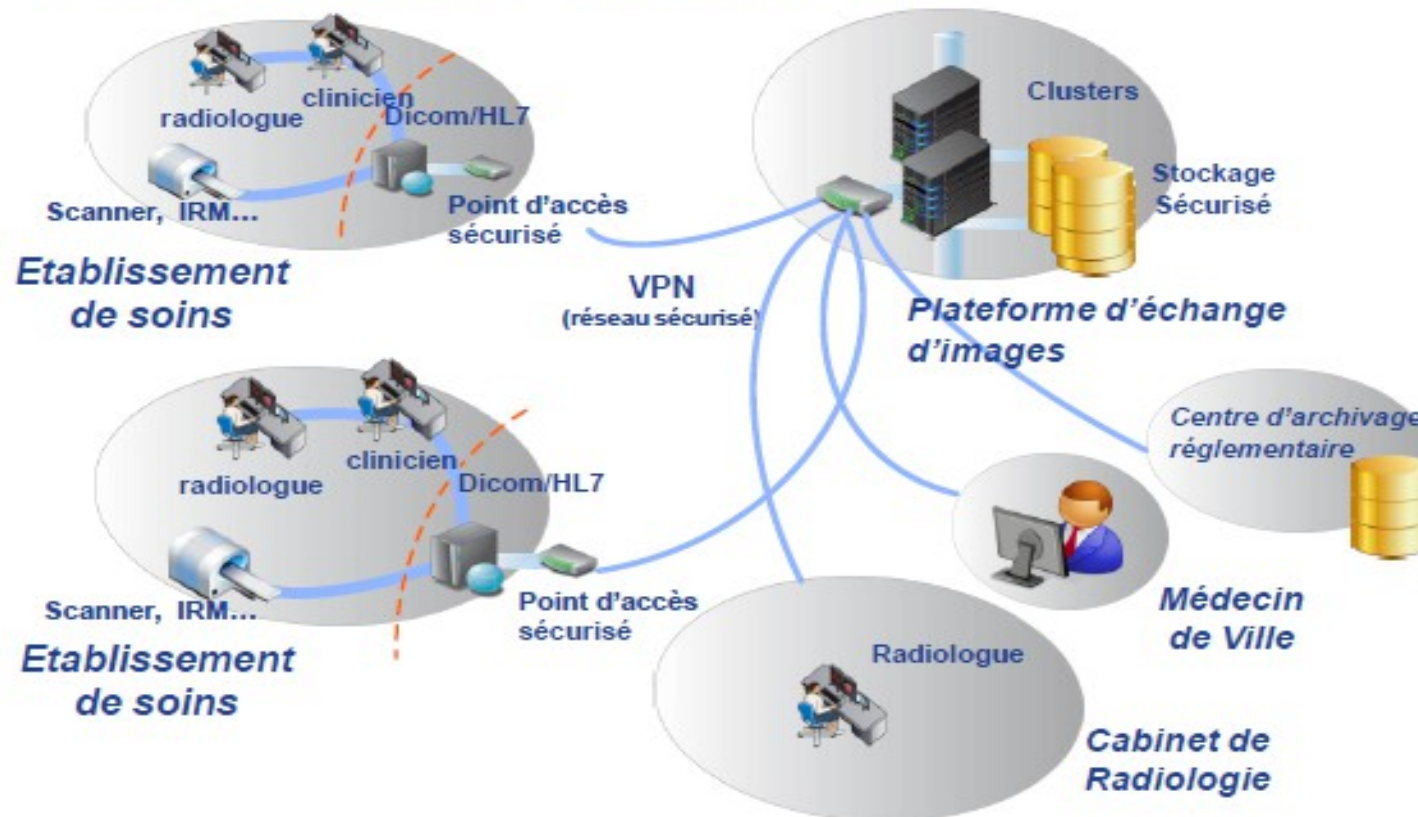
Maintenant : le numérique





# Architecture fonctionnelle d'un PACS

Système de gestion des images médicales avec des fonctions d'archivage, de stockage, permettant le partage entre professionnels d'images et de données cliniques ou biologiques. Intégré au système informatique hospitalier.

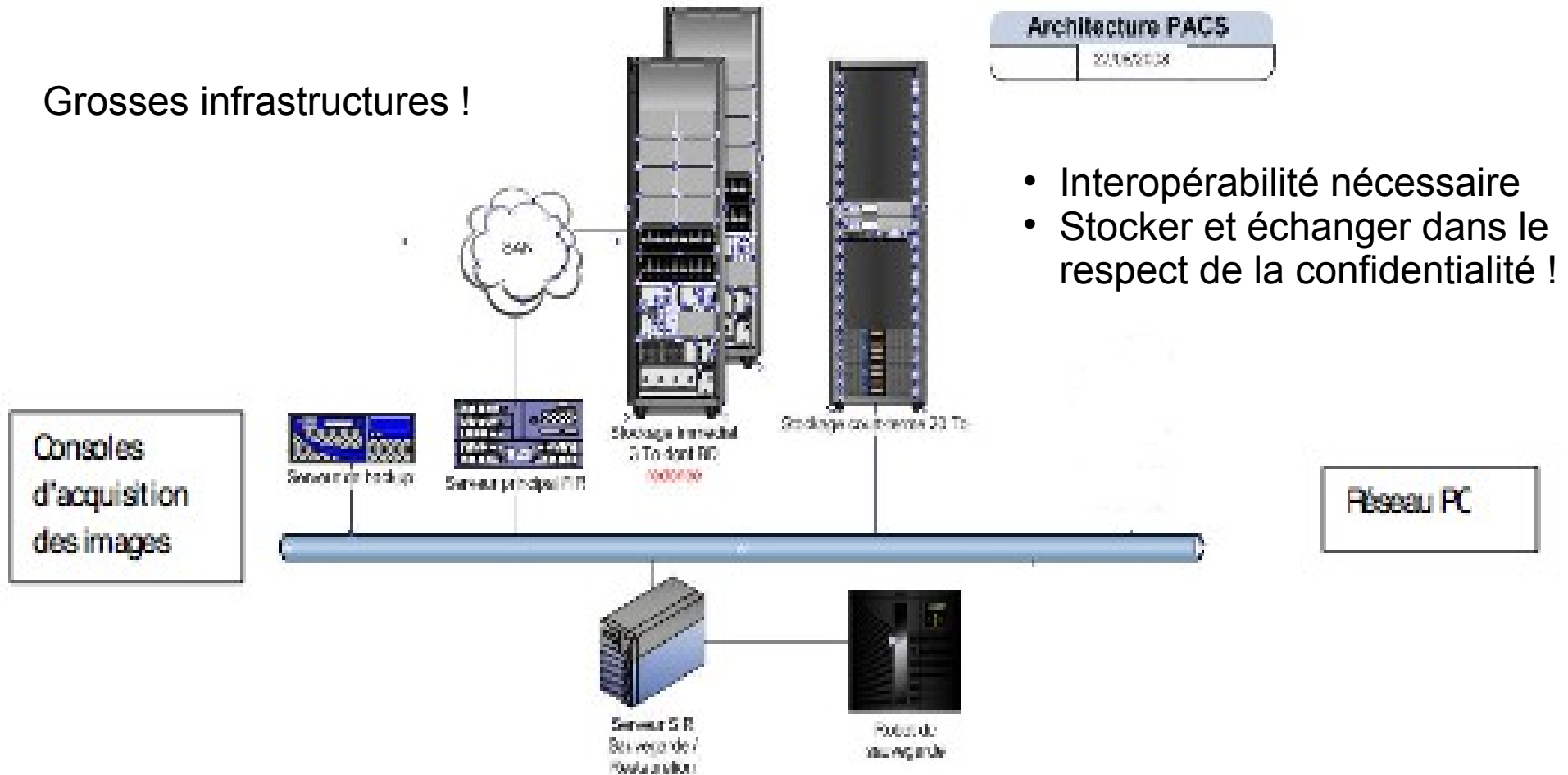


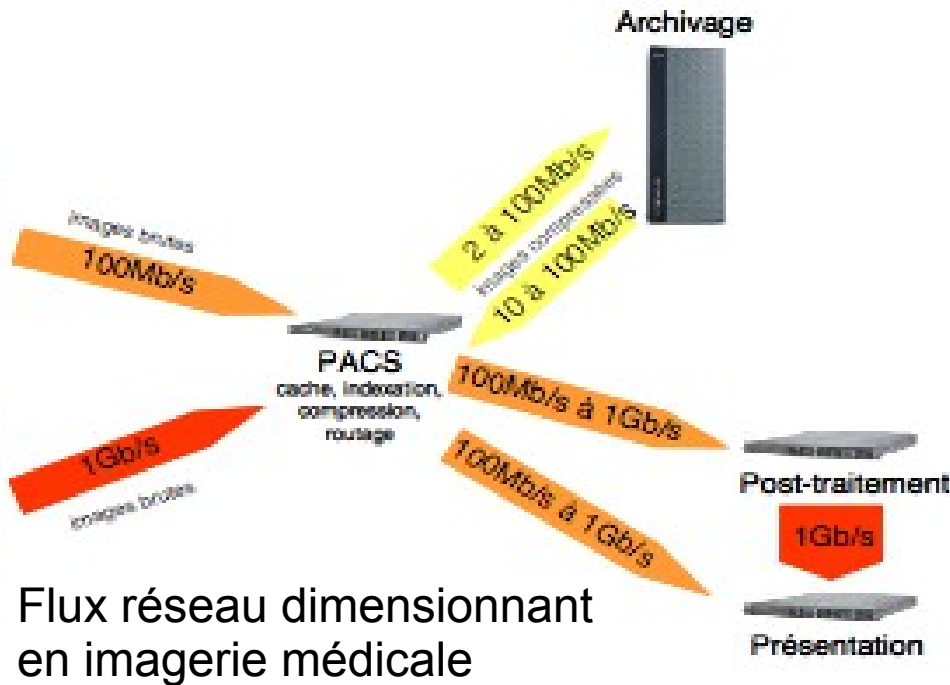
Permet de centraliser les images provenant de tous les systèmes d'imagerie  
=> diagnostic plus rapide et thérapies plus sûres (partage d'informations à distance)  
=> économique (suppression des examens redondants, des déplacements, archivage)

# Les structures

- Programme « **Hôpital numérique** » pour le développement et la modernisation des SIH
- Programme « **Territoire des soins numériques** » cible partage et échange d'informations pour prise en charge coordonnée tout au long du parcours de soins dans et hors hôpital.

Grosses infrastructures !

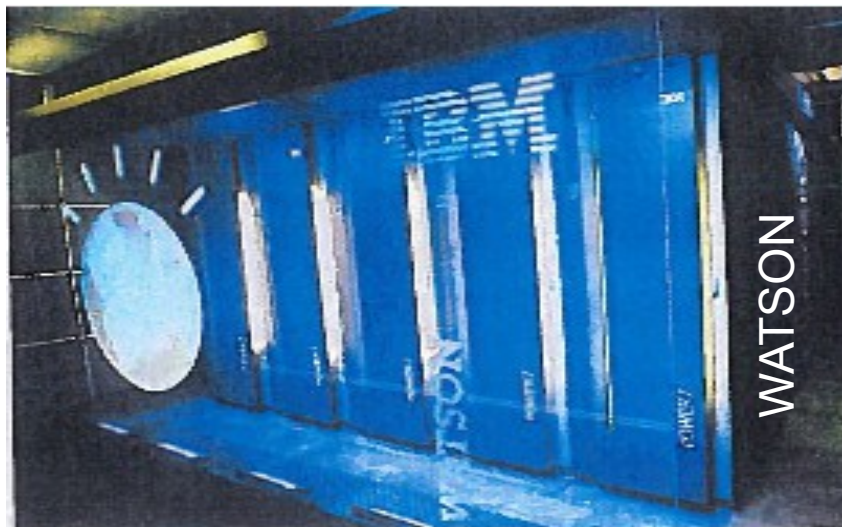




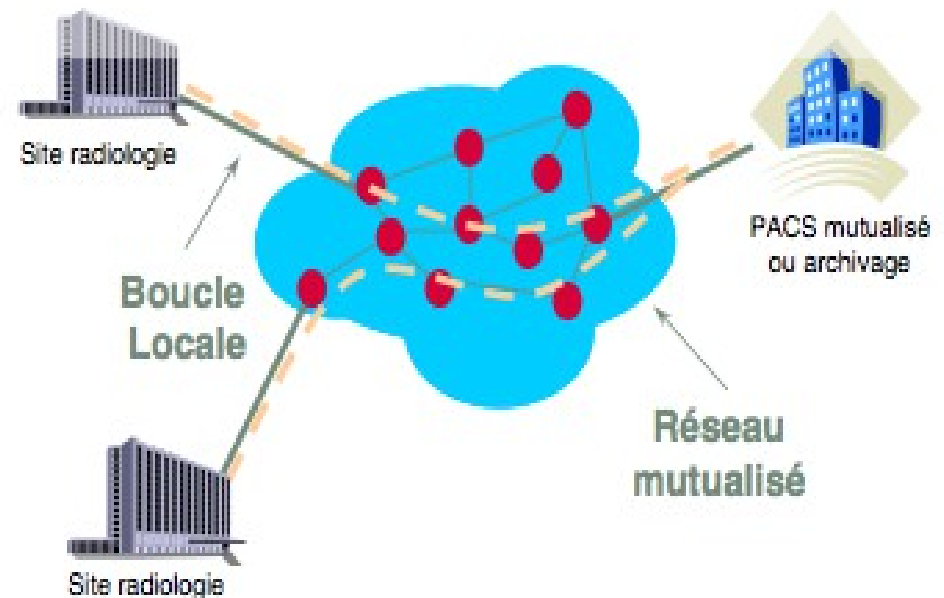
Hébergeur de données de santé  
 - **CHEOPS Technology** avec 11 agences en France  
 - **Groupe SIGMA** avec 3 data centers à Nantes,  
 agréés par le ministère de la santé

**Evolution vers le Cloud computing**  
 Moins couteux  
 En principe neutre...

mais les données sont bien stockées dans des centres matériels, presque tous aux USA !

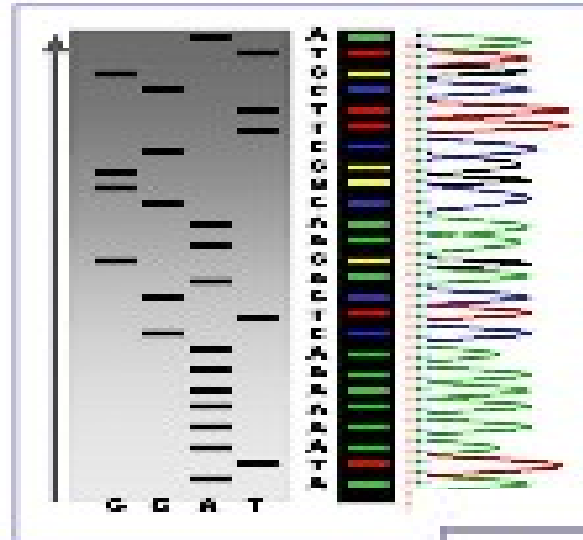
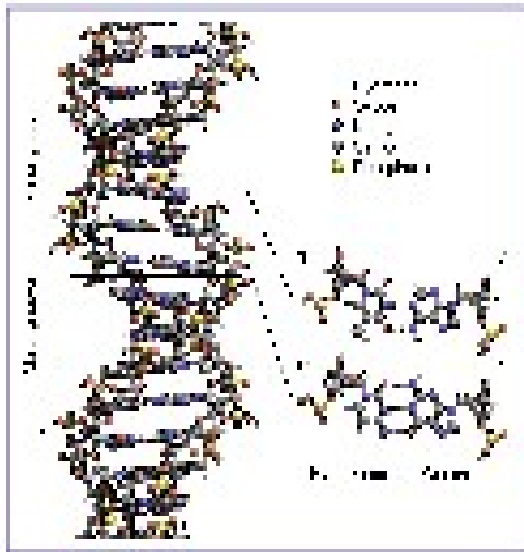


Watson analyse en quelques secondes 1,5 million de dossiers médicaux.





# Séquençage du génome : la première application Big Data

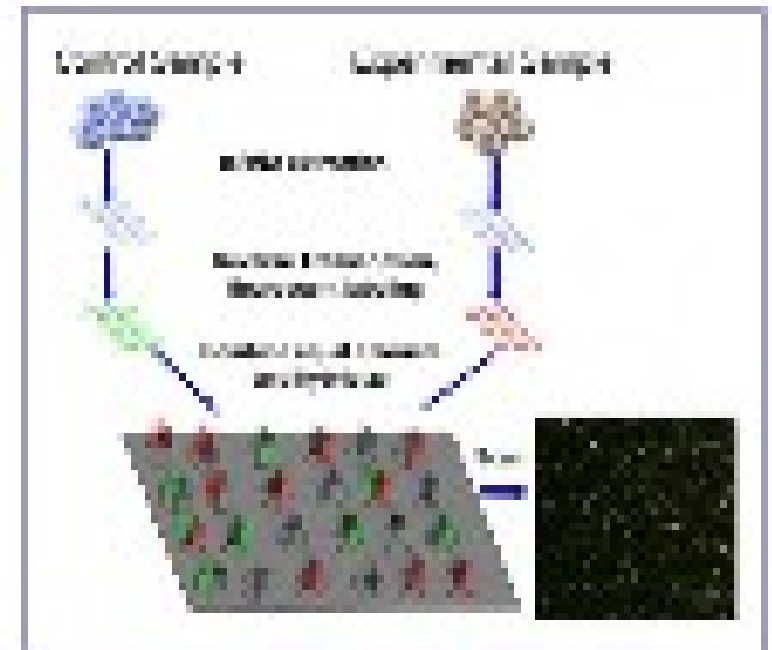


Épeler la chaîne d'ADN et déterminer l'enchaînement des nucléotides le long de la séquence

## Explosion des données

- les puces à ADN
- la collecte via les réseaux
- séquençage massif et parallèle
- statistique et modèles

**Les risques d'exploitation :**  
eugénisme, assurances et mutuelles...



## Outil puissant ! Effrayant ? Mais un réel intérêt !

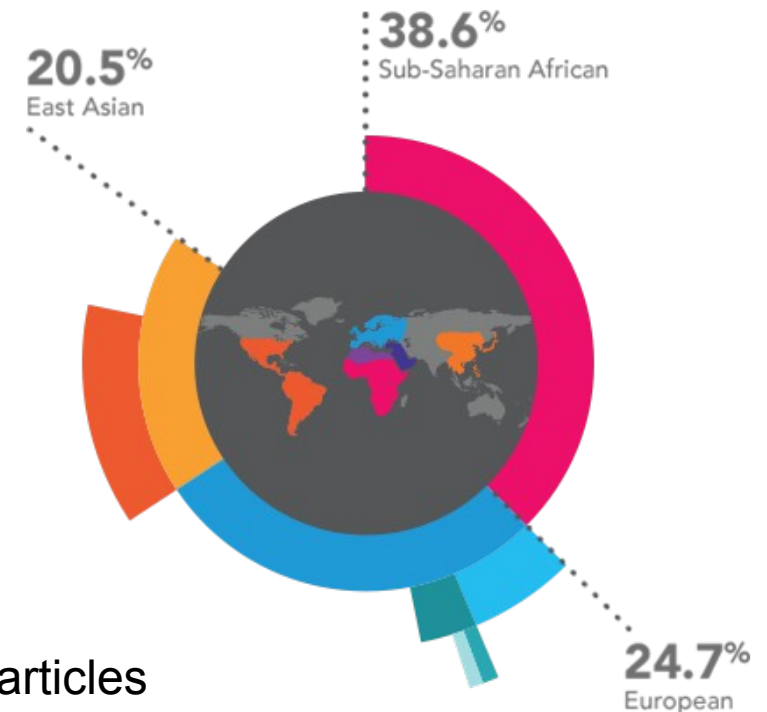
Aux USA, des start-up ont lancé des applications permettant d'obtenir une carte ADN sur mesure pour obtenir les informations relatives aux maladies qui vous guettent !

- => juste un peu de salive
- => informations sur l'origine

**Genomic Data Sharing Policy** : résultat de plusieurs directives sur le partage des données établies depuis 2003, pour promouvoir le partage des données génomiques. Accès déjà fourni à plus de 2400 chercheurs de 41 pays.

**Le projet BrainSCANr** , exploitant plusieurs millions d'articles scientifiques, a permis d'identifier un lien entre des maladies et certaines parties du cerveau.

**Un programme d'intelligence artificielle d'IBM, baptisé Watson** a absorbé des millions de pages d'articles spécialisés et de rapports cliniques. On en attend une meilleure connaissance de la maladie et des pistes pour mieux la prédire et donc la prévenir.



# Les données issues des objets connectés

La quête de l'immortalité !

**Google** : nanoparticules pour détecter et soigner le cancer..

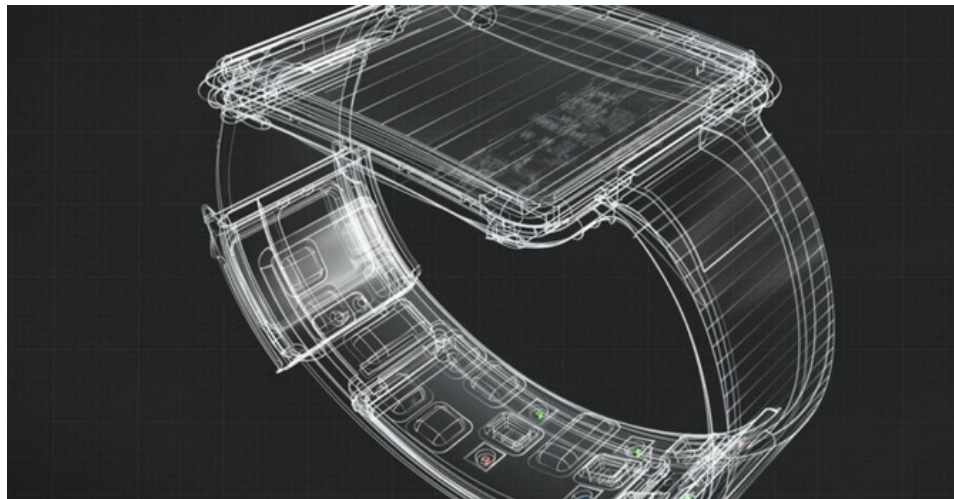
**Samsung** : Capteur de rythme cardiaque dans le Galaxy S5...

Simband en cours de développement associé à une plateforme big data (Samsung Architecture Multimodal Interactions) pour concentrer les données et les vendre !

« *Les données seront disponibles pour les partenaires qui veulent créer des applications et services approuvés* »

**Apple**: montre qui collecte en permanence des données ou iphone => Health kit

Interdit l'exploitation publicitaire aux développeurs (ex produits amincissement pour personnes en surpoids) mais garde tous les droits y compris de vendre ces données à des assurances privées



# Données personnelles : une marchandise comme une autre !

## Développement de l'open data : une demande de l'industrie

- La plus grande banque mondiale de données de santé est en France : SNIIR-AM (système d'informations inter régimes de l'assurance maladie)
- Celtipharm : utilisation des feuilles de soin pour études sur prise de médicaments et situations à risques (1,3 Mds de données par an !)
- Think Thank Orange Healthcare (Healthcare Data Institute) : la chaîne de valeur est différente selon les pays.
- La CNIL a autorisé AXA en 2010 à utiliser les données de soins pharmaceutiques de ses adhérents pour optimiser leurs contrats.

## Maladies chroniques et Télémédecine

- => questions techniques (déploiement)
- => questions éthiques (confidentialité)
- => questions humaines (changement de culture du personnel de santé, responsabilité des patients)

**De nombreuses formations** : parcours « Numérique et Santé » à l'ISEP (école d'ingénieurs)

**De nombreuses réflexions !**

Big data peut traiter des données exprimées dans toutes les langues. Il peut aussi exploiter d'autres données contenues dans les dessins vectoriels, les images matricielles (en pixels), la vidéo, le son, les logs, les données spatiales (géolocalisation), les sites web, les blogs, les mails, les échanges sur les réseaux sociaux (Facebook, Twitter, LinkedIn...), les éléments de biométrie, etc.

Hadoop : ensemble logiciel structurel de licence libre, géré par la fondation Apache, il utilise des algorithmes de type MapReduce 2.0 qui, après avoir découpé les milliards d'enregistrements en blocs de même taille, répartit l'exécution du traitement « parallèle distribué » sur des serveurs distants, puis enchaîne les résultats reçus (concaténation) avant d'adresser le résultat sous une forme visuelle (data visualisation).

Note du Centre de Recherche de l'Ecole des Officiers de la Gendarmerie Nationale